Stochastic Derivative-free Optimization Using a Trust Region Framework

Jeffrey Larson
KTH Automatic Control Group
FEBRUARY 5, 2016

The Problem

We want to solve:

$$\underset{x \in \mathbb{R}^n}{\text{minimize}} \, f(x)$$

when $\nabla f(x)$ is unavailable and we only have access to noise-corrupted function evaluations $\bar{f}(x)$.

The Problem

We want to solve:

$$\operatorname*{minimize}_{x \in \mathbb{R}^n} f(x)$$

when $\nabla f(x)$ is unavailable and we only have access to noise-corrupted function evaluations $\bar{f}(x)$.

Such noise may be deterministic (e.g., from iterative methods) or stochastic (e.g., from a Monte-Carlo process).

The Problem

We want to solve:

$$\underset{x \in \mathbb{R}^n}{\text{minimize}}\, f(x)$$

when $\nabla f(x)$ is unavailable and we only have access to noise-corrupted function evaluations $\bar{f}(x)$.

Such noise may be deterministic (e.g., from iterative methods) or stochastic (e.g., from a Monte-Carlo process).

Model-based methods are one of the most popular methods when ∇f is unavailable, and the only recourse when noise is deterministic.

The Problem

We analyze the convergence of our method in the stochastic case:

$$\bar{f}(x) = f(x) + \epsilon,$$

where ϵ is identically distributed with mean 0 and variance $\sigma^2 < \infty$.

The Problem

We analyze the convergence of our method in the stochastic case:

$$\bar{f}(x) = f(x) + \epsilon,$$

where ϵ is identically distributed with mean 0 and variance $\sigma^2 < \infty$.

This is equivalent to solving:

$$\underset{x}{\text{minimize}} \ \mathbb{E}\left[\bar{f}(x)\right].$$

Example

Stochastic Approximation

Iterates usually have the form:

$$x_{k+1} = x_k + a_k G(x_k),$$

where

- $G(x_k)$ is a cheap, unbiased estimate for $\nabla f(x_k)$
- a_k is a sequence of step sizes

Stochastic Approximation

Iterates usually have the form:

$$x_{k+1} = x_k + a_k G(x_k),$$

where

- $G(x_k)$ is a cheap, unbiased estimate for $\nabla f(x_k)$
- a_k is a sequence of step sizes (specified by the user) satisfying:

$$\sum_{k=1}^{\infty} a_k = \infty \qquad \lim_{k \to \infty} a_k = 0$$

Stochastic Approximation

Iterates usually have the form:

$$x_{k+1} = x_k + a_k G(x_k),$$

where

- $G(x_k)$ is a cheap, unbiased estimate for $\nabla f(x_k)$
- a_k is a sequence of step sizes (specified by the user) satisfying:

$$\sum_{k=1}^{\infty} a_k = \infty \qquad \lim_{k \to \infty} a_k = 0$$

Algorithm performance depends significantly on sequence a_k.

Response Surface Methodology

- Developed by the experimental design community.

- Build models using a fixed pattern of points, for example, cubic, spherical, or orthogonal designs among many others.

- Finding the design that constructs response surfaces approximating the function without requiring excessive function evaluations can be difficult for problems where the user has no prior expertise.

Modifications to Existing Methods

Take a favorite method and repeatedly evaluate the function at points of interest.

- Stochastic approximation modified by Dupuis, Simha (1991)
- Response surface methods modified by Chang et al. (2012)
- UOBYQA modified by Deng, Ferris (2006)
- Nelder-Mead modified by Tomick et al. (1995)
- DIRECT modified by Deng, Ferris (2007)

Modifications to Existing Methods

Take a favorite method and repeatedly evaluate the function at points of interest.

- Stochastic approximation modified by Dupuis, Simha (1991)
- Response surface methods modified by Chang et al. (2012)
- UOBYQA modified by Deng, Ferris (2006)
- Nelder-Mead modified by Tomick et al. (1995)
- DIRECT modified by Deng, Ferris (2007)

There are two downsides to such an approach:

1. Repeated sampling only provides information about the noise ϵ, not f.
2. If the noise is deterministic, no information is gained.

Overview

We therefore desire a method that

1. Adjusts the step size as it progresses
2. Does not use a fixed design of points
3. Does not repeatedly sample points

Overview

We therefore desire a method that

1. Adjusts the step size as it progresses

2. Does not use a fixed design of points

3. Does not repeatedly sample points

We aim to extend the convergence results of Conn, Scheinberg, and Vicente for trust-region methods to the stochastic case.

Similarly, we'd like the class of possible models to be general.

κ-**fully Linear**

Definition

If $f \in LC$ and $\exists$ a vector $\kappa = (\kappa_{ef}, \kappa_{eg})$ of positive constants such that

- the error between the gradient of the model and the gradient of the function satisfies

$$\|\nabla f(y) - \nabla m(y)\| \leq \kappa_{eg}\Delta \ \ \forall y \in B(x; \Delta),$$

- the error between the model and the function satisfies

$$|f(y) - m(y)| \leq \kappa_{ef}\Delta^2 \ \ \forall y \in B(x; \Delta),$$

we say the model is κ-*fully linear* on $B(x; \Delta)$.

α-**probabilistically** κ-**fully Linear**

Definition

Let $\kappa = (\kappa_{ef}, \kappa_{eg})$ be a given vector of constants, and let $\alpha \in (0, 1)$. Let $B \subset \mathbb{R}^n$ be given. A random model m_k generated at the kth iteration of an algorithm is α-probabilistically κ-fully linear on B if

$$P\left(m_k \text{ is a } \kappa\text{-fully linear model of } f \text{ on } B \big| \mathcal{F}_{k-1}\right) \geq \alpha,$$

where $\mathcal{F}_{k-1}$ denotes the realizations of all the random events for the first $k - 1$ iterations.

Regression Models can be α-probabilistically κ-fully Linear

Theorem

For a given $x \in \mathbb{R}^n$, $\Delta > 0$, $\alpha \in (0,1)$,

- *$Y \subset B(x; \Delta)$ is strongly Λ-poised,*
- *The noise present in $\bar{f}$ is i.i.d. with mean 0, variance $\sigma^2 < \infty$,*
- *$|Y| \geq C/\Delta^4$,*

Then there exist constants $\kappa = (\kappa_{ef}, \kappa_{eg})$ (independent of Δ and Y) such that the linear model m regressing Y is α-probabilistically κ-fully linear on $B(x; \Delta)$.

One Last Part

For our analysis, we need estimates of $f(x^k)$ and $f(x^k + s^k)$ that are slightly different than those provided by the model functions.

Let F_k^0 and F_k^s denote the sequence of estimates of $f(x^k)$ and $f(x^k + s^k)$.

We need to be able to construct estimates satisfying

$$\mathbb{P}\left[\left|F_k^0 - f(x^k)\right| > \epsilon \min\left\{\Delta_k, \Delta_k^2\right\} \big| \mathcal{F}_{k-1}\right] < \theta$$
$$\text{and } \mathbb{P}\left[\left|F_k^s - f(x^k + s^k)\right| > \epsilon \min\left\{\Delta_k, \Delta_k^2\right\} \Big| \mathcal{F}_{k-1}\right] < \theta,$$

for any $\epsilon > 0$ and $\theta > 0$.

Algorithm 1: A trust-region algorithm to minimize a stochastic function

---

Pick $0 < \gamma_{dec} < 1 < \gamma_{inc}$, $0 < \eta$, $\beta < 1$, $0 < \Delta_0$, and $\alpha \in (0,1)$. Set $k = 0$;

Start

Build a α_k-probabilistically κ-fully linear model m_k on $B(x^k; \Delta_k)$ for some $\alpha_k \geq \alpha$;

Compute $s^k = \arg \min\limits_{s: \|x^k - s\| < \Delta_k} m_k(s)$;

if $m_k(s^k) - m_k(x^k + s^k) \geq \beta \Delta_k$ **then**

    Calculate $\rho_k = \dfrac{F_k^0 - F_k^s}{m_k(x^k) - m_k(x^k + s^k)}$;

    if $\rho_k \geq \eta$ **then**

        Calculate $x^{k+1} = x^k + s^k$; $\Delta_{k+1} = \gamma_{inc}\Delta_k$;

    else

        $x^{k+1} = x^k$; $\Delta_{k+1} = \gamma_{dec}\Delta_k$;

    end

else

    $x^{k+1} = x^k$; $\Delta_{k+1} = \gamma_{dec}\Delta_k$;

end

$k = k + 1$ and go to **Start**;

Convergence

If the function f, noise ϵ, and various algorithmic constants satisfy some assumptions, we prove that our algorithm converges almost surely to a first-order stationary point of f.

Convergence

Assumption

The additive noise ϵ observed when computing $\bar{f}$ is independent and identically distributed with mean zero and bounded variance σ^2.

Assumption

On some set $\Omega \subseteq \mathbb{R}^n$ containing all iterates visited by the algorithm,

- *f is Lipschitz continuous*
- *∇f is Lipschitz continuous*
- *f has bounded level sets*

Convergence

Assumption

The constants $\alpha \in (0, 1)$, $\gamma_{dec} \in (0, 1)$, and $\gamma_{inc} > 1$ satisfy

$$\alpha \geq \max\left\{ \frac{1}{2}, 1 - \frac{\frac{\gamma_{inc}-1}{\gamma_{inc}}}{4\left[\frac{\gamma_{inc}-1}{2\gamma_{inc}} + \frac{1-\gamma_{dec}}{\gamma_{dec}}\right]} \right\},$$

where

- α *is the lower bound on the probability of having a κ-fully linear model,*
- $\gamma_{dec} \in (0, 1)$ *is the factor by which we decrease the trust region radius,*
- $\gamma_{inc} > 1$ *is the factor by which the trust radius is increased.*

Convergence

Assumption

The constants $\alpha \in (0,1)$, $\gamma_{dec} \in (0,1)$, and $\gamma_{inc} > 1$ satisfy

$$\alpha \geq \max \left\{ \frac{1}{2}, 1 - \frac{\frac{\gamma_{inc}-1}{\gamma_{inc}}}{4 \left[\frac{\gamma_{inc}-1}{2\gamma_{inc}} + \frac{1-\gamma_{dec}}{\gamma_{dec}} \right]} \right\},$$

where

- α *is the lower bound on the probability of having a κ-fully linear model,*
- $\gamma_{dec} \in (0,1)$ *is the factor by which we decrease the trust region radius,*
- $\gamma_{inc} > 1$ *is the factor by which the trust radius is increased.*

If $\gamma_{inc} = 2$ and $\gamma_{dec} = 0.5 \implies \alpha \geq 0.9$.
If $\gamma_{inc} = 2$ and $\gamma_{dec} = 0.9 \implies \alpha \geq 0.65$.

Proof Outline

Theorem

If the above assumptions are satisfied, our algorithm converges almost surely to a first-order stationary point of f.

- Show the sequence of trust region radii $\Delta_k \to 0$ almost surely.

Proof Outline

Theorem

If the above assumptions are satisfied, our algorithm converges almost surely to a first-order stationary point of f.

- Show the sequence of trust region radii $\Delta_k \to 0$ almost surely.
- Show if Δ_k ever falls below some constant multiple of the model gradient, $\Delta_{k+1} > \Delta_k$ with high probability.

Proof Outline

Theorem

If the above assumptions are satisfied, our algorithm converges almost surely to a first-order stationary point of f.

- Show the sequence of trust region radii $\Delta_k \to 0$ almost surely.
- Show if Δ_k ever falls below some constant multiple of the model gradient, $\Delta_{k+1} > \Delta_k$ with high probability.
- Lastly, show that, the sequence of ratios

$$\left\{ \frac{\|\nabla f(x_k)\|}{\Delta_k} \right\}$$

  is bounded above by a nonnegative supermartingale. Since every nonnegative supermartingale converges almost surely, and $\Delta_k \to 0$ almost surely, this implies $\|\nabla f(x_k)\| \to 0$ almost surely.

Problem Set

53 problems of the form:

$$f(x) = \sum_{i=1}^{m} \left[(1 + \sigma)F_i^s(x)\right]^2,$$

where $\sigma \sim U[-0.1, 0.1]$.

Problem Set

53 problems of the form:

$$f(x) = \sum_{i=1}^{m} \left[(1 + \sigma) F_i^s(x) \right]^2,$$

where $\sigma \sim U[-0.1, 0.1]$.

If S is the set of solvers to be compared on a suite of problems P, let $t_{p,s}$ be the number of iterates required for solver $s \in S$ on a problem $p \in P$ to find a function value satisfying:

$$f(x) - f_L \leq \tau \left(f(x_0) - f_L \right),$$

where f_L is the best function value achieved by any $s \in S$.
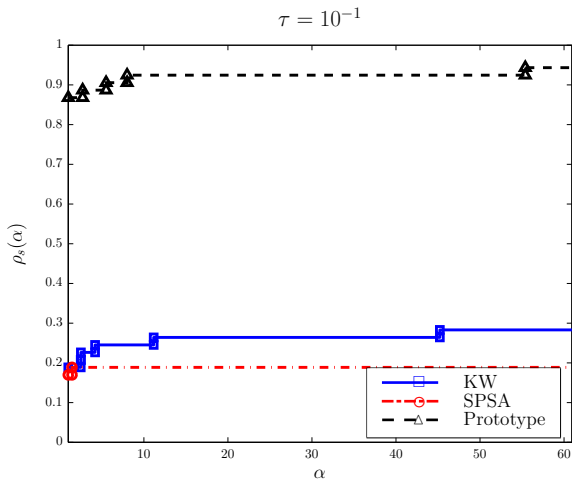
Performance Profile

Then the performance profile of a solver $s \in S$ is the following fraction:

$$\rho_s(\alpha) = \frac{1}{|P|} \left| \left\{ p \in P : \frac{t_{p,s}}{\min \left\{ t_{p,s} : s \in S \right\}} \leq \alpha \right\} \right|$$

Performance Profile

Then the performance profile of a solver $s \in S$ is the following fraction:
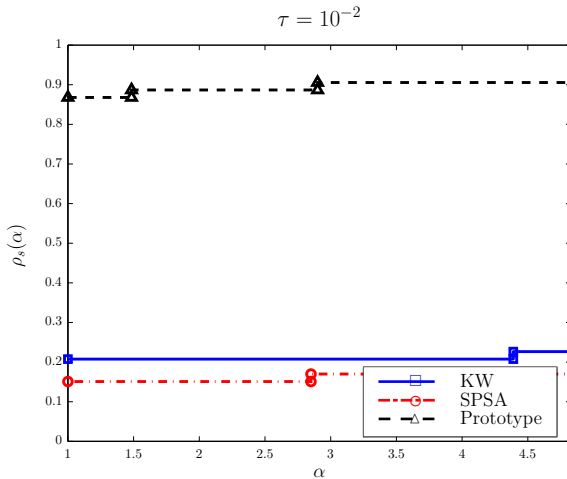
$$\rho_s(\alpha) = \frac{1}{|P|} \left| \left\{ p \in P : \frac{t_{p,s}}{\min \{t_{p,s} : s \in S\}} \leq \alpha \right\} \right|$$

- Note that we are using the true function value f, not the observed $\bar{f}$.
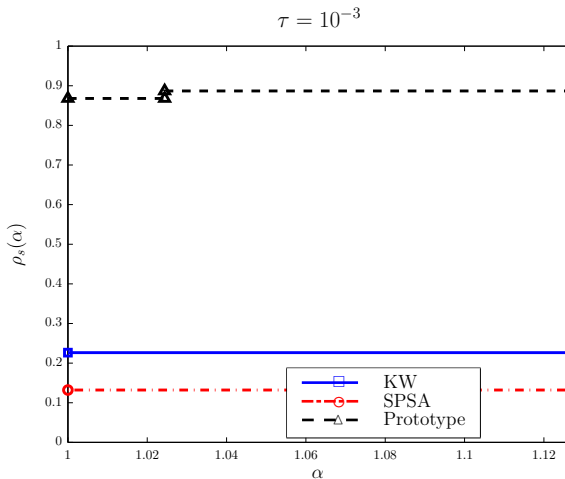- Since the noise is stochastic, each solver is run 10 times per problem.

Performance Profile

$\tau = 10^{-1}$

Performance Profile

$$\tau = 10^{-2}$$

Performance Profile

Current Work

- Generalizing results to ensure a practical algorithm converges.
  - For example, not requiring α-probabilistically κ-fully linear models every iteration.

Current Work

- Generalizing results to ensure a practical algorithm converges.
  - For example, not requiring α-probabilistically κ-fully linear models every iteration.

- Smartly constructing α-probabilistically κ-fully linear models.